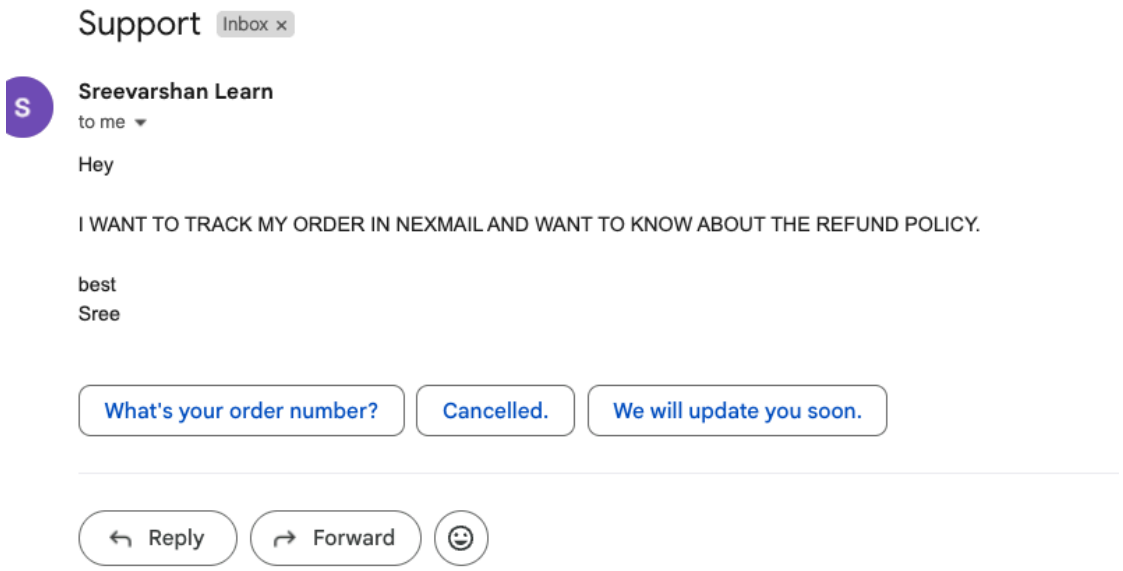
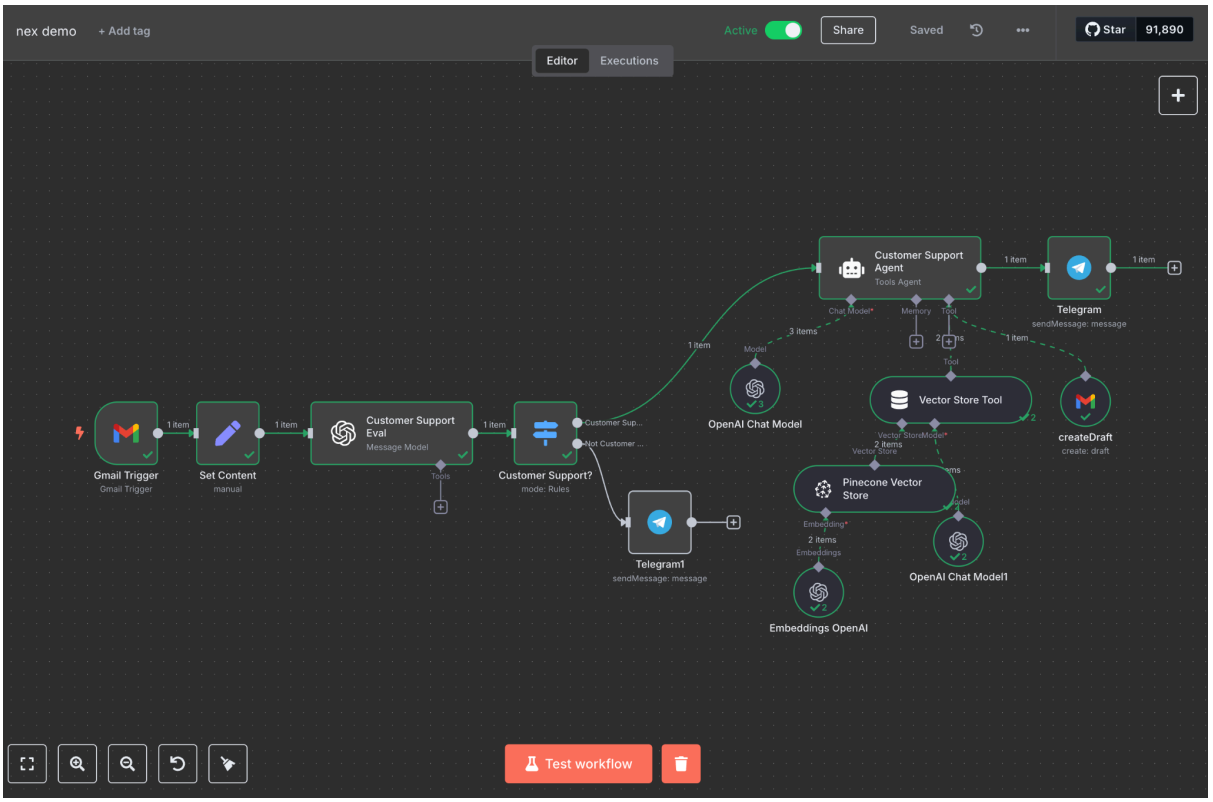
NexMail AI Customer Support Automation

BREAKDOWN

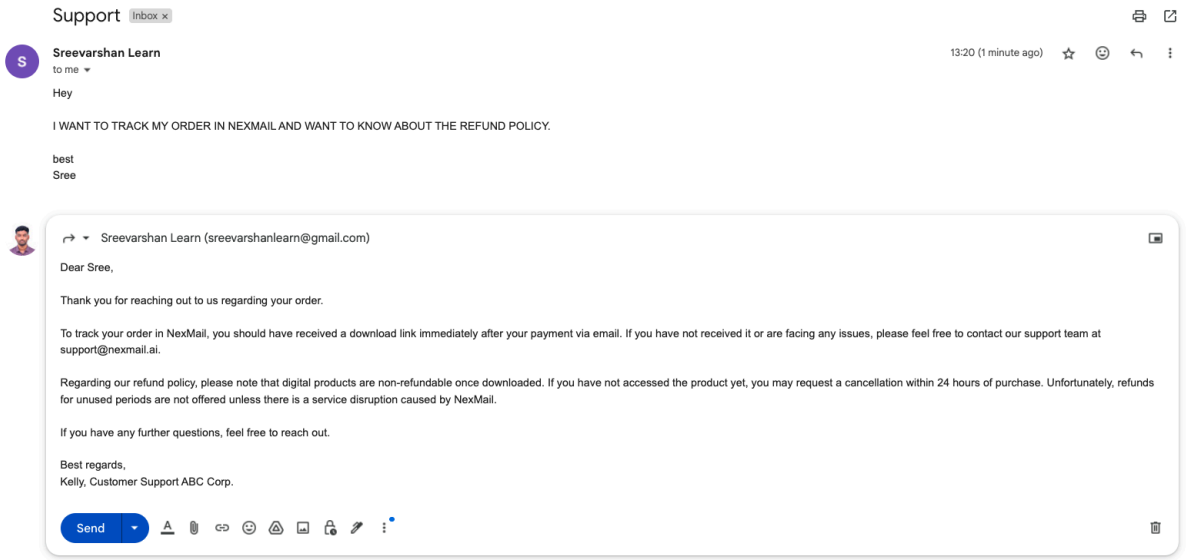
E-mail received from the customer



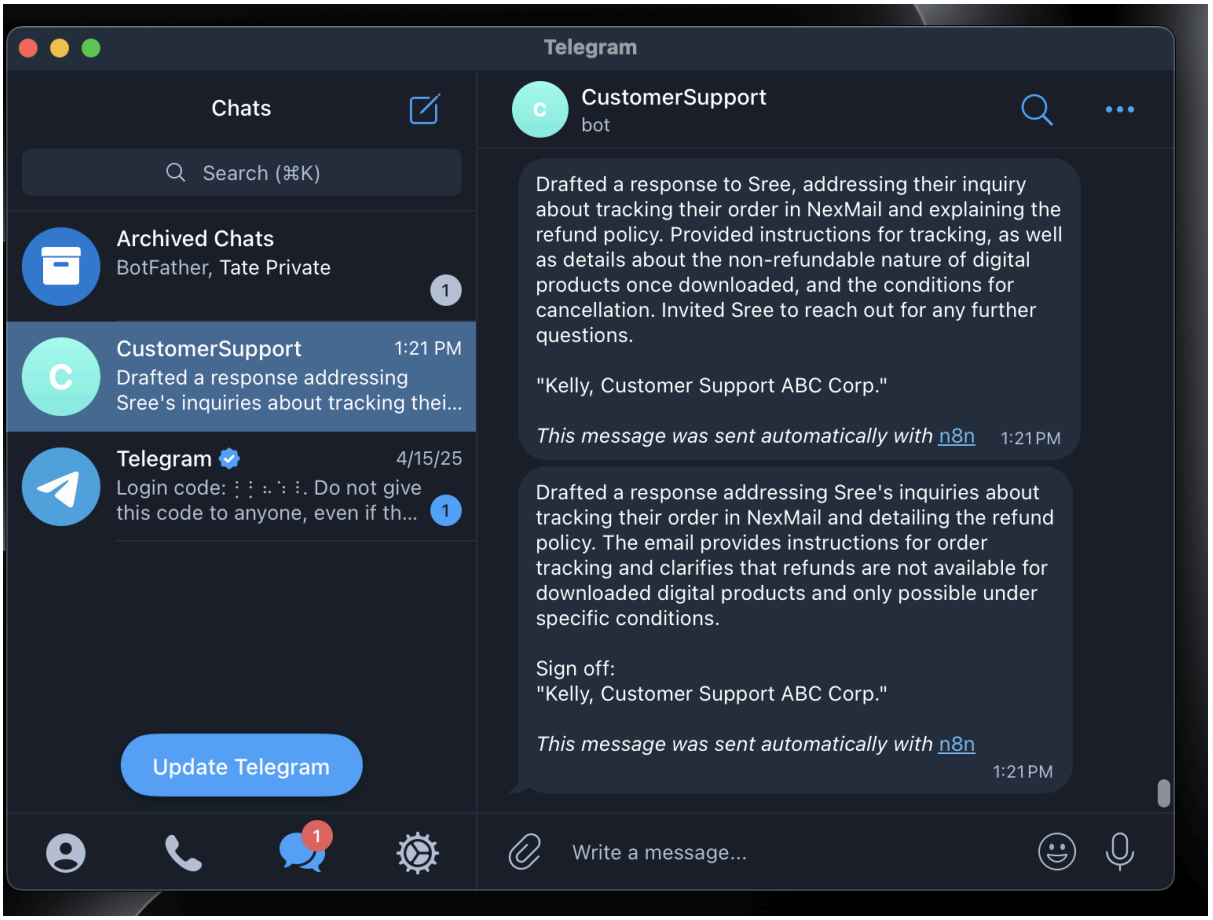
Active workflow from n8n working automatically



It has generated a draft message by referencing the database using Pinecone.



It triggers an automatic telegram message from the bot named CustomerSupport



NexMail AI Customer Support Automation

Technical Implementation Report

Prepared by: Sreevarshan Sathiyamurthy
Date: May 13, 2025
Contact: sathiyamurthy.sr@northeastern.edu

Executive Summary

This report outlines the technical design and implementation of NexMail's AI-powered customer support automation system. Built using N8N, OpenAI's large language models, and a Pinecone vector database, the system delivers precise, context-aware, and instant responses to customer inquiries.

In comparison to previously implemented N8N-based support methods — which largely relied on rule-based logic, keyword filtering, and pre-defined templates — this solution dramatically improves query understanding, response quality, and operational scalability. It establishes a new benchmark in support automation by unifying intelligent query processing with seamless multi-channel delivery.

System Architecture Overview

The solution architecture integrates multiple intelligent components to form a seamless end-to-end automation flow:

- **Email Trigger Workflow**
  Captures and initiates workflows based on new customer inquiries via Gmail.

- **Content Analysis Pipeline**
  Extracts, interprets, and classifies incoming messages using custom logic and AI models.

- **Vector Database Integration**
  Uses semantic search to retrieve support content with high contextual relevance.

- **AI Response Generation**
  Generates dynamic, empathetic replies using OpenAI's GPT models.

- **Multi-Channel Delivery**
  Sends responses via email and optionally through Telegram, with tracking and performance logging.

Technologies Implemented

Core Components

- **N8N Workflow Automation**
  Orchestrates the data pipeline, connects APIs, and triggers responses.

- **OpenAI GPT Model (Chat + Embeddings)**
  Powers both understanding of user intent and generation of natural language replies.

- **Pinecone Vector Database**
  Provides a high-dimensional search engine for semantic retrieval of support content.

- **Gmail API Integration**
  Enables real-time monitoring and extraction of customer emails.

Technical Specifications

Feature	Description
Vector Embedding Model	`text-embedding-3-small` – optimized for semantic similarity
Content Classifier	Custom model to detect and categorize support topics
Infrastructure	AWS Lambda & S3 in `us-east-1` (serverless, event-driven)
Vector Dimensions	1536-dimensional embeddings for fine-grained relevance
Deployment Method	Stateless, serverless design with dense vector retrieval
Security Compliance	TLS encryption, OAuth 2.0 for Gmail, API key vaulting for Open Pinecone

---

Implementation Workflow

1. Email Detection & Triggering

- Gmail trigger node monitors the support inbox.
- When an email is received, metadata and content are extracted and passed to the pipeline.

2. Content Analysis & Classification

- Subject, body, and metadata are structured using the `Set Content` node.
- A classifier model determines whether the query is support-related and identifies the topic.

3. Knowledge Retrieval

- A semantic search is performed against the Pinecone vector store using the customer's message as the query.
- Top results are selected based on cosine similarity (≥ 0.7 threshold).
- Retrieved content includes support policies, FAQs, and recent documentation.

4. Response Generation

- GPT-powered Chat node synthesizes the retrieved content into a concise, personalized, and brand-aligned response.
- A template engine formats the reply consistently with NexMail's tone and guidelines.

5. Multi-Channel Delivery

- Responses are delivered via Gmail's API directly to the customer.
- Optional Telegram delivery is triggered for cross-platform visibility.
- Metadata (response time, classification confidence, delivery status) is logged for analysis.

Performance Metrics & Benefits

Category	AI-Powered System	Improvement Over Prior Methods
Speed of Resolution	Responses are generated in under 5 minutes with contextual understanding.	~50% faster than template-based automation with delays
Query Understanding	Capable of parsing multi-intent, sentiment-driven, and ambiguous queries.	~2x better recognition than rule/keyword-based flows
Information Accuracy	Semantic retrieval ensures relevant, policy-aligned content every time.	~40% more accurate than FAQ lookup or static message flows
Adaptability	Adapts dynamically to query phrasing and tone.	Far more flexible than rigid decision-tree logic

Channel Flexibility	Easily integrates email, Telegram, and future channels.	No need for separate flows per platform
Maintenance Overhead	Only requires updating vector database or policies (no logic re-wiring).	~60% lower upkeep vs node-by-node logic-based updates
Workflow Depth	Integrates detection → classification → retrieval → generation → delivery in one seamless flow.	~2x deeper coverage than modular, segmented automations
Scalability	Scales to thousands of queries per day on serverless infra.	Easily 3x the capacity of current N8N queue/thread models

Business Impact

- **Customer Experience**
  Delivers fluid, well-structured replies that feel human — outperforming template-based systems that often feel robotic or irrelevant.

- **Support Team Optimization**
  Routine and repeat queries are handled end-to-end by the automation, allowing human agents to focus on edge cases and relationship-building tasks.

- **Consistency in Communication**
  Unlike manually updated N8N reply nodes, the AI draws from a single, centralized policy base — eliminating contradictory or outdated responses.

- **Reduced Operational Overhead**
  Policy updates don't require editing workflows. Uploading updated knowledge content or documentation keeps the system aligned without engineering effort.

- **Scalable Architecture**
  Easily expandable to support other regions, languages, or even live chat integrations — all without duplicating complex automation logic.

Conclusion

- Respond with speed and accuracy.
- Scale without inflating staffing costs.
- Ensure consistency across all customer interactions.
- Maintain a low-maintenance, extensible infrastructure.